Forensic voice comparison using sub-band cepstral distances as features: A first attempt with vowels from 306 Japanese speakers under channel mismatch conditions

Yuko Kinoshita[1], Takashi Osanai[2], Frantz Clermont[3]
[1] College of Arts and Social Science/Asia and the Pacific, The Australian National University
[2]National Research Institute of Police Science, Japan
[3]J.P. French Associates Forensic Lab., England
osanai@nrips.go.jp; yuko.kinoshita@anu.edu.au; dr.fclermont@gmail.com

Abstract

This study presents the latter part of an exploratory study of the potential of sub-band parametric cepstral distance (PCD) as an alternative forensic voice comparison (FVC) feature to formants and cepstral coefficients. Using 5 Japanese vowels produced by 306 male Japanese speakers, we conducted LR-based FVC experiments under a channel mismatch condition, with sub-bands selected in reference to the expected formant locations. Combining 3 sub-band PCDs from F1, F2, and F3 ranges, sub-band PCDs outperformed the full-band PCDs in speaker classification, demonstrating their promise as an automatically extractable, robust, and linguistically interpretable acoustic feature for FVC.

Index Terms: Sub-band cepstral distance, likelihood ratio, forensic voice comparison, channel mismatch, Japanese vowels

1. Introduction

Both speech and voice recognition systems are now part of our daily lives, and yet forensic voice comparison (FVC hereafter) is still no easy task. One of the reasons for this is the lack of control over the data. The speech samples in FVC, especially those from crime scenes, are often short and contain considerable background noise. The scarcity of data means insufficient data for modelling of speakers. Poor recording quality compromises the accuracy of acoustic feature extraction. As speakers are modelled based on those acoustic features, this also contributes to poor quality of the speaker models.

Also, the speech samples to be compared are most often recorded under very different circumstances. The speakers may be in very different emotional states, speak in different styles, and also be recorded on different devices via different transmission channels. These factors, which are unrelated to speaker characteristics, can amplify the acoustic differences between two samples, contributing to difficulties in producing strong likelihood ratios (LRs) in support of the same-speaker hypothesis, even where the speakers are indeed the same.

What the analyst can do to improve the situation is limited. We may be able to improve the elicitation and recording process of the known (or suspect) speaker, but crime scene recordings are largely out of our control.

Over the years, much research has been done on the impact of channel mismatch (e.g. [1-3]), and various techniques have been proposed to compensate for channel mismatch (e.g. [4-6]). However, such techniques all appear to require building a channel characteristics model. Crime scene recordings are often very short, so the recording in question may not contain sufficient information to build a reliable channel characteristics model. Also, mobile transmission characteristics change continuously, as the compression rate and methods change in response to network conditions [7, 8]. This makes the alternative approach, i.e. retrospectively 'matching' the conditions by putting a non-telephone recording (such as a police interview) through a mobile codec or a telephone network, less attractive. Further, various social network platforms now offer voice call options. It is thus increasingly unlikely for analysts to have access to full information on the processing applied to the speech sample in question.

These issues suggest that the most practical way forward in FVC is to search for features which are robust under forensically realistic conditions: less affected by external factors, and reliably measurable even with poor quality of recordings. This led us to the sub-band parametric cepstral distance (PCD), an approach initially proposed in [9]. PCD extracts the difference between two cepstral shapes within user-defined frequency boundaries. Its potential has been discussed in two studies: [10] examined within- and between-speaker variability of PCD, using landline telephone speech recordings from 297 Japanese speakers. Another study [11] made small scale observations on the F-ratio of sub-band PCDs using mobile and microphone recordings. The results from both studies were encouraging.

This motivated us to embark on the current project: an examination of the potential of sub-band PCD as an FVC feature using a large dataset. As the first step, we examined the behavior of sub-band PCDs in detail with respect to their F-ratios and verification rates in different sub-band ranges, using a database of Japanese vowels elicited from 306 speakers [12]. This database permits us to examine the forensically significant effect of channel mismatch, as it was recorded simultaneously via two channels: microphone and mobile phone transmission. The results of the initial experiments were promising; they suggested that sub-band PCD is relatable to articulatory gestures in similar ways to formants. This brings two advantages specific to forensic application: firstly, the results can be explained in court to non-experts in a relatively less abstract way; secondly any unusual results can be detected and reexamined in relation to articulatory and phonetic characteristics, more easily than full-band cepstra. They also found that speaker verification based on sub-band PCDs degrades less under a channel mismatch condition compared to that based on full-band PCDs, presumably because sub-band PCDs can exclude frequency ranges unrelated to speaker information.

This paper thus continues to examine the potential of sub-band PCDs as a speaker classification feature by selecting sub-

band ranges based on vowel formant frequencies, and conducting LR-based voice comparison experiments under a channel mismatch condition.

2. Data and procedures

2.1. Database, speakers, and speech materials

This study used the same data as [12]: 306 adult male speakers from the NRIPS database [13]. They are native speakers of Japanese, aged from 18 to 76 years. They had widely varied dialectal background, but dialectal variations appear not to affect vowel formants much in modern Japanese [14]. Thus the dialectal variation is unlikely to have contributed to greater between-speaker variability here. All speakers were recorded on two occasions, 2 to 3 months apart. They performed the same recording tasks twice at each recording session, and the whole process was recorded simultaneously through 2 channels: direct microphone (Ch1), and via a mobile phone network (Ch3). This study focuses on the cross-channel comparisons.

Read-out (C)V syllables were used as the speech samples: that is, Japanese 5 vowel phonemes, /a/, /e/, /i/, /o/ and /ɯ/, preceded by selected consonantal environment: ∅ (no consonant), /k/, /s/, /t/, /h/, /r/, /g/, /z/, /d/, /b/, and /p/. The phonemes /n/, /m/, /y/, and /w/ were excluded from analysis this time to facilitate reliable automatic segmentation. These are highly controlled elicitation, not spontaneous. However, [14] reports relatively small vowel reduction in running speech in Japanese. Therefore, we regard the current data as acceptable for this exploratory work.

Japanese *kana* syllabary writing system maintains the distinction between the pairs ぢ /di/ – じ /zi/ and づ /dɯ/ – ず /zɯ/, but they are phonetically identical, both realized as [dʑi] and [dzɯ]. Consequently, we have the vowel data in 10 different phonological contexts for /i/ and /ɯ/, and 11 for /a/, /e/, and /o/.

2.2. Segmentation and full-band LPCC extraction

The target syllables were automatically segmented into a preceding consonant and a vowel based on their power and F0. The sound files were down-sampled from 44.1 kHz to 8kHz, and full-band LPCCs were extracted from the selected vowel sections (order 14, Hamming window, window length 25ms, time-step 5ms). The LPCCs were averaged across the vowel duration, and further averaged across different phonological contexts for each vowel. As result, we obtained LPCCs for 5 vowels, 2 recording sessions, 2 repeats, and 2 recording channels for each speaker.

2.3. Parametric cepstral distance (PCD) calculation

The parametric cepstral distance (PCD) described in [9] affords selection of any sub-band range directly from full-band LPCCs. Its formulation is summarised below in Eq. (1), where $D^2\left(\overline{\mathbf{C}}_i, \overline{\mathbf{C}}_j, \omega_1, \omega_2\right)$ represents the Euclidean distance between any pair of full-band LPCCs $(\overline{\mathbf{C}}_i, \overline{\mathbf{C}}_j)$ for a given sub-band range. Note that the full-band LPCCs are index-weighted by the matrix $\mathbf{K}$ to emphasise spectral slope differences, and then weighted by the matrix $\mathbf{W}(\omega_1, \omega_2)$ to focus on any sub-band range selectable by its lower and upper limits ω_1 and ω_2. For $\omega_1 = 0$ and $\omega_2 = \pi$, Eq. (1) simply reduces to the familiar Euclidean distance between any pair of (index-weighted) full-band LPCCs.

$$D^2\left(\overline{\mathbf{C}}_i, \overline{\mathbf{C}}_j, \omega_1, \omega_2\right) = \left(\overline{\mathbf{C}}_i - \overline{\mathbf{C}}_j\right)^T \cdot \mathbf{K}^T \cdot \mathbf{W}(\omega_1, \omega_2) \cdot \mathbf{K} \cdot (\overline{\mathbf{C}}_i - \overline{\mathbf{C}}_j) \equiv \text{PCD between } \overline{\mathbf{C}}_i \text{ and } \overline{\mathbf{C}}_j \quad (1)$$

where:
$i, j \equiv$ speaker-session index
$\overline{\mathbf{C}}_i \equiv$ mean LPCC for i^{th} speaker across all tokens
$\overline{\mathbf{C}}_j \equiv$ mean LPCC for j^{th} speaker across all tokens

$$\mathbf{K} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \boldsymbol{k} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \dots & \mathbf{0} & \boldsymbol{M} \end{bmatrix} \equiv \text{index-weighting matrix}$$

$\mathbf{W}(\omega_1, \omega_2) \equiv$ band-selective matrix (see [9])
$\boldsymbol{M} \equiv$ LPCC order = 14
$\omega_1 \equiv$ lower limit of sub-band selected within $[0, \pi]$
$\omega_2 \equiv$ upper limit of sub-band selected within $[0, \pi]$

Sub-band PCD has the capacity to limit the analysis to the user-defined frequency regions, allowing us to exclude frequency regions that are unhelpful in assessing speaker identity. In the first part of this project, we found that the F-ratios tend to be higher in the frequency regions where we expect to find formants [12]. Thus, this time we select the sub-band frequency ranges referring to the formant measurements made in [15]. For each vowel, the mean ± 1 standard deviation of the first three formants were sought. The frequency ranges which contain the above values to the nearest 100Hz were defined as the target sub-band ranges for this study. These ranges are referred to as subF1, subF2 and subF3 hereafter.

Table 1. *Target sub-band ranges for each vowel (Hz).*

	subF1		**subF2**		**subF3**	
	from	to	from	to	from	to
/a/	600	800	1200	1600	2300	2800
/e/	300	600	1800	2200	2500	2900
/i/	200	400	1900	2400	2600	3100
/o/	300	600	1000	1300	2300	2700
/ ɯ /	200	400	1300	1800	2200	2700

2.4. Comparisons

With 306 speakers recorded 4 times (2 non-contemporaneous occasions, twice per sessions), we had 1224 patterns of same-speaker (SS) pairs and 3373320 patterns of different-speaker (DS) pairs. All comparisons were made in cross-channel conditions, i.e. between direct microphone recording (Ch1) and mobile phone network recording (Ch3).

2.5. Modelling and LR calculation

For LR calculation in linguistics-based FVC research, MVKD proposed by [16] has been a popular choice. It is, however, inappropriate to put PCDs through the MVKD formula, as a PCD is already a distance measure between two sets of information. The PCDs from the SS pairs and those from the DS pairs represent within- and between-speaker variations of the distance between two cepstra in the regions of the user-selected frequency ranges.

The relatively large data size in this study suggests that general population is reasonably well represented by the current data. However, examination of the PCD distributions revealed that the 1224 comparisons for SS were not sufficient to produce a smooth distribution, and direct derivation of LR

from it will result in some arbitrary fluctuations of LRs. The distributions need to be modelled.

To find an appropriate model, we tested the fit of four different distributions: normal, gamma, Weibull, and log normal, with the PCDs from different vowel and band combinations. Both SS and DS comparisons were evaluated for their fit to those 4 distributions based on Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). Although AIC and BIC measures fitting slightly differently [17], both measures selected an identical model as the best fit in all vowel and band range combinations in this study. Table 2 presents the counts of each model which produced the lowest AIC and BIC. SS and DS indicate the comparison types. The maximum score for each cell is 5, as we tested for all 5 vowels. Gamma clearly outperformed the rest.

Table 2: *Number of instances each model was selected as the best fit ()*

	norm		**gamma**		**weibull**		**lnorm**	
	SS	DS	SS	DS	SS	DS	SS	DS
Full	0	0	4	5	0	0	1	0
subF1	0	0	4	2	1	3	0	0
subF2	0	0	4	3	1	2	0	0
subF3	0	0	5	4	0	1	0	0
total	**0**	**0**	**17**	**14**	**2**	**6**	**1**	**0**

Based on this result, we fitted gamma distribution to the distributions of PCDs. Here, we added another type of PCD: sum of the PCDS from subF1, subF2 and subF3. This equates to the sum of area differences in three sub-band regions obtained from a pair of cepstra. Five vowels, 2 comparison types, and 3 sub-band ranges + full-band + summed PCD, resulted in 50 distributions. All were modelled with gamma distributions defined as below:

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta} \qquad (2)$$

We tested if lack of independence between testing and distribution modelling data has any effect by modelling the distributions with some speakers removed. We repeated 100 times removing a different set of 6 speakers each time, but no meaningful effect was found, as expected from the data size. Thus LRs were calculated by: 1) pooling PCDs separately for SS and DS comparisons and modelling their distributions, 2) deriving probabilities of the testing pairs to be belonging to the 2 different distributions, applying them to the models. The obtained LRs were then converted to Log_{10}LR (LLR) and calibrated. Cllr [18] was also calculated.

3. Results and Discussion

3.1. LLRs

In this section, we add another feature combination, summed LLR: the sum of 3 LLRs obtained from subF1, subF2, and subF3. Summing potentially correlated LLRs such as these risks introducing inaccuracy. However, the correlation among LLRs turned out to be very low, as seen in Table 3. The strongest correlation coefficient found was 0.188 (between subF1 and subF2 of /o/ vowel), indicating that correlations is unlikely to distort the results significantly.

Figure 1 presents the mean LLRs for each vowel and band range selection. It reveals that SS and DS comparisons are separated well at the theoretical threshold, LLR 0, across all vowels and band ranges. The vowel which produces strongest LRs — i.e. appearing at the furthest positions from 0 on both directions — is /ɯ/, closely followed by /i/. /a/ and /o/ appear to produce weaker LLRs. Comparing subF1, subF2 and subF3, we can see that subF1 is of limited use. The speaker information seems to be most richly carried in subF2.

Table 3: *Correlation between LLRs (Pearson's r)*

	SS comparisons			**DS comparisons**		
	F1-F2	*F1-F3*	*F2-F3*	*F1-F2*	*F1-F3*	*F2-F3*
a	0.072	-0.022	0.174	0.013	0.024	0.098
e	0.032	0.087	0.169	0.003	0.019	0.161
i	0.055	0.049	0.046	0.027	-0.003	0.131
o	0.188	0.128	0.045	0.045	0.109	-0.052
ɯ	0.102	0.121	0.102	0.040	0.053	0.055

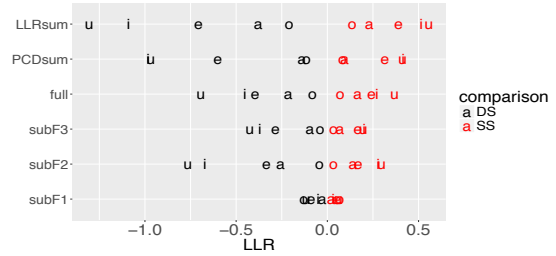

Figure 1: *Mean LLRs for each vowel and band range*

The results from the first part of this study [12] and the theoretical nature of the sub-band PCD predict sub-band PCDs (such as sum of PCD and sum of LLR) to outperform the full-band PCDs. Figure 1 shows that this is indeed to be the case; sum of PCD, and sum of LLR outperformed from full-band, confirming utility of band-selective analyses.

3.2. Verification rate and Cllr

Next, we observe the rates of successful speaker verification at threshold LLR 0. We focus our observation in this section on the comparison between full-band, sum of PCDs and sum of LLRs. With all vowels, the LLRs supported the correct hypothesis well above chance level, /i/ and /ɯ/ reaching over 80% for DS, which is a strong result for a single vowel. For all vowels but /a/, the sub-band based approach constantly outperformed full-band. Even for /a/, sum of LLR performed better than full-band. Here too the high vowels /i/ and /ɯ/ performed better than other vowels.

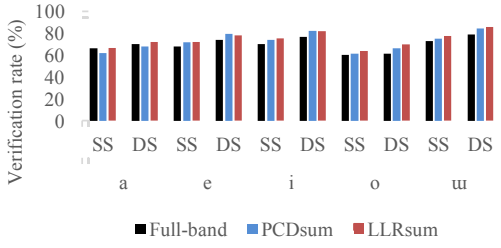

Figure 2: *Successful classification rate at LLR 0*

Cllr is the cost metric that evaluates the quality of the classification system [18]. We used half of the SS and DS comparisons for training the calibration, and the rest to

examine Cllr. Cllr can be decomposed to the verification cost (Cllr_min) and the calibration cost (Cllr_cal). For ease of interpretation, the decomposed components are presented in Figure 3. The results are presented separately for each vowel, /a/ to /ɯ/ from the bottom to the top. The categories "Full-band", "3_area", and "3_LLR" indicate full-band PCD, sum of PCD, and sum of 3 sub-band LLR.

Across all vowels, the scores for Cllr_cal (in dark blue) were very low, indicating that the system was already well calibrated, and the classification errors were largely caused by the PCD's discriminant capacity. The calibration results confirmed this; Cllrs did not improve with calibration, as seen in Figure 4, which presents the comparison of pre and post calibration Cllr, pooled across all band types in a violin plot.

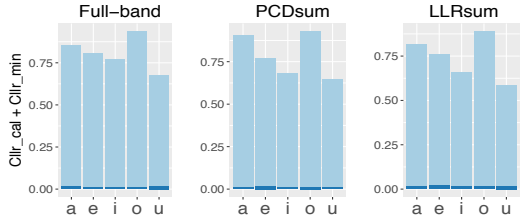

Figure 3: *Pre-calibration Cllr_min and Cllr_cal*

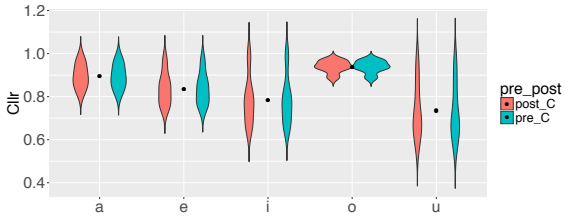

Figure 4: *Pre-and post-calibration difference in Cllr*

4. Conclusion

This paper further examined the behavior of sub-band PCDs. We selected 3 sub-bands for each vowel based on their known formant frequency ranges, and examined their performance under a channel mismatch condition. The results showed that individual sub-band PCDs were not as powerful as the full-band PCDs but once combined, they outperformed the full-band PCDs as predicted. Also LLRs produced from PCDs were found to be extremely well calibrated.

The examinations presented in [12] and here support our proposition of sub-band PCD being a potentially useful FVC feature. Most results were predictable from existing phonetic knowledge, suggesting sub-band PCD to be a feature that is automatically extractable and more readily interpretable – a desirable quality for evidence presentation in court.

As future tasks, performance comparison to the existing approaches is critical, especially to formant-based FVC. We also plan to do further work on optimal sub-band ranges, and the effect of sample data size and speech style, and different approaches to the LR calculation.

5. Acknowledgements

The work presented here was partly supported by JSPS KAKENHI Grant Number JP18H01671, JP25350488.

6. References

[1] H. J. Künzel, "Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies," *Forensic Linguistics* vol. 8, pp. 80-99, 2001.

[2] C. Byrne and P. Foulkes, "The 'mobile phone effect' on vowel formants," *International Journal of Speech Language and the Law,* vol. 11, pp. 83-102, 2004.

[3] C. Zhang, G. S. Morrison, E. Enzinger, and F. Ochoa, "Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – Female voices," *Speech Communication,* vol. 55, pp. 796-813, 7// 2013.

[4] A. A. Garcia and R. J. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, 1999, pp. 325-328.

[5] D. A. Reynolds, "Channel robust speaker verification via feature mapping," ed: I E E E, 2003, pp. II-53-6.

[6] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005, pp. I/629-I/632 Vol. 1.

[7] B. J. Guillemin and C. Watson, "Impact of the GSM mobile phone network on the speech signal: some preliminary findings," *International Journal of Speech, Language & the Law,* vol. 15, 2008.

[8] E. A. Alzqhoul, B. B. Nair, and B. J. Guillemin, "Impact of dynamic rate coding aspects of mobile phone networks on forensic voice comparison," *Science & Justice,* vol. 55, pp. 363-374, 2015.

[9] F. Clermont and P. Mokhtari, "Frequency-band specification in cepstral distance computation," in *The 5th Australian International Conference on Speech Science & Technology* 1994, pp. 354-359.

[10] M. Khodai-Joopari, F. Clermont, and M. Barlow, "Speaker variability on a continuum of spectral sub-bands from 297-speakers' non-contemporaneous cepstra of Japanese vowels," in *The 10th Australian International Conference on Speech Science and Technology*, Sydney, 2004, pp. 504-509.

[11] F. Clermont, Y. Kinoshita, and O. Takashi, "Sub-band cepstral variability within and between speakers under microphone and mobile conditions: A preliminary investigation," in *The 16th Australasian International Conference on Speech Science & Technology*, Sydney, 2016.

[12] T. Osanai, Y. Kinoshita, and F. Clermont, "Exploring sub-band cepstral distances for more robust speaker classification," presented at the 17th Speech Science and Technology Conference (SST2018), Sydney, 2018.

[13] H. Makinae, T. Osanai, T. Kamada, and M. Tanimoto, "Construction and preliminary analysis of a large-scale bone-conducted speech database," *IEICE technical report,* vol. Speech 107, pp. 97–102, 2007.

[14] 奥田浩三, "発話スタイルの変動に頑健な音響モデル構築法に関する研究," 大阪市立大学, 2005.

[15] Y. Kinoshita, "Testing Realistic Forensic Speaker Identification In Japanese: A Likelihood Ratio Based Approach Using Formants," PhD, Linguistics, The Australian National University, Canberra, 2001.

[16] C. Aitken, G.G. and D. Lucy, "Evaluation of trace evidence in the form of multivariate data," *Applied Statistics,* vol. 53, pp. 109--122, 2004.

[17] S. I. Vrieze, "Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)," *Psychological Methods,* vol. 17, pp. 228-243, 2012.

[18] D. A. van Leeuwen and N. Brümmer, "An introduction to application - Independent evaluation of speaker recognition system," in *Speaker Classification*. vol. 1, C. Müller, Ed., ed Berlin: Springer, 2007, pp. 330--353.